'Birds of a feather' fail together

Citation for published version:
Calabrese, R, Andreeva, G & Ansell, J 2019, ''Birds of a feather' fail together: Exploring the nature of dependency in SME defaults', Risk analysis, vol. 39, no. 1, pp. 71-84. https://doi.org/10.1111/risa.12862

Digital Object Identifier (DOI):
10.1111/risa.12862

Published In:
Risk analysis

‘Birds of a feather’ fail together: exploring the nature of dependency in SME defaults

Abstract

This paper studies the effects of incorporating the interdependence among London small business defaults into risk analysis framework using the data just before the financial crisis. We propose an extension from standard scoring models to take into account the spatial dimensions and the demographic characteristics of SMEs, such as legal form, industry sector and number of employees. We estimate spatial probit models using different distance matrices based only on the spatial location or on an interaction between spatial locations and demographic characteristics. We find that the interdependence or contagion component defined on spatial and demographic characteristics is significant and that it improves the ability to predict defaults of non-start-ups in London. Furthermore, including contagion effects among SMEs alters the parameter estimates of risk determinants. The approach can be extended to other risk analysis applications where spatial risk may incorporate correlation based on other aspects.

Keywords: spatial probit model, proximity matrix, small business, scoring model.

1 Introduction

Support and development of a small business is fundamental to a nation's economic growth and is seen as an important function of any government. Access to credit forms a vital pre-requisite for business growth, and it is not surprising that following the recent financial crisis banks have been urged to increase lending to Small and Medium Sized Enterprises (SMEs). Nevertheless, any increases in lending have been modest due to inherently risky nature of credit and a justified tendency for prudential lending. Risk analysis forms an integral part of decision-making support in lending and involves a variety of qualitative and quantitative approaches, summarised in Kazemi and Mosleh (2012). One of the quantitative decision support paradigms is credit scoring, it is widely used in consumer credit (Thomas et al., 2002), and it is this type of decision support system that this paper uses as a starting point in analysis of SME failures.

Several authors (Berger and Udell, 2002; Grunert and Norden, 2012) noted that risk analysis of SMEs is particularly difficult because of insufficient or unverifiable information (e.g. non-audited financial statements). Besides, failures or defaults are often clustered together or linked to each other, and this is usually not incorporated into the existing methodology of credit scoring. On the one hand, there are studies in spatial econometrics that concentrate on modelling spatial dependencies, on the other hand, there is research on credit risk and credit scoring, with very limited intersection between these two areas of knowledge. The situation is similar to that of multi-criteria spatial decision support systems as described in Ferretti and Montibeller (2016).

This paper explores an approach that brings together the theory of spatial modelling and decision-making framework of credit scoring by incorporating the dependency between small business failures/defaults into a credit scoring model. The necessity to account for such a dependency has emerged during the recent credit crisis, which has demonstrated the so-called credit contagion or the interdependence of borrowers that amplifies the spread of defaults. There have been proposals for different models for corporate defaults (Battiston et al., 2012; Delli Gatti et al., 2009; Egloff et al., 2007 and Neu et al., 2004), yet again for SMEs this line of investigation is limited, and the existing studies have focussed on spatial dependence. Barro and Basso (2010) proposed a model that incorporated a network of interdependent businesses and used a spatial proximity to represent business connections. Their study considered the dependence between geographical regions, it was based on simulation and did not involve testing on real data. To investigate spatial dependence between the companies Fernandes and Artes (2016) applied an ordinary kriging model to defaults of Brazilian SMEs and found that it improves the credit scoring model. Spatial dependence was also established to be present for the growth of Brazilian SMEs (Cravo et al., 2015).

The current paper builds on the above studies and explores the nature of the dependence between SMEs defaults that can be incorporated into the decision making process of financial institutions. The main objective and contribution of this paper consists in investigating how dependence between SMEs arising from demographic characteristics can be overlaid with spatial proximity in risk analysis of failures. The analysis is based on a sample of London SMEs that are at least three years old. We do not find the evidence for spatial dependence in our data, this may happen as the London area is affected by similar economic conditions. In contrast, there is a statistically significant interdependence when the spatial proximity is interacted with company demographics (industry, legal form, etc.). The novelty consists in extending the spatial framework to incorporate a different type of dependency by using distance/ proximity measures that are common in data mining, thus viewing dependence from a completely different angle.

To the best of our knowledge this is the first study of this type, and the findings will inform the inter-disciplinary research, including spatial econometricians, small business experts and risk analysts, in general. The approach can be extended to a number of spatial risk applications, e.g. in health or ecology, where spatial risk may be aggravated by the correlation arising from other aspects.

The rest of the paper is structured as follows: Section 2 is literature review. It summaries the existing research on modelling the performance and failure risk of SMEs; and models of credit contagion and spatial dependence. Section 3 presents the methodology of spatial modelling and proximity matrices. This is followed by data description, empirical results and impact on decision-making in Section 4, whilst the last Section 5 concludes with discussion and suggestions for further research.

2 Literature review

2.1 Modelling SMEs performance and risk of failure

Modelling a failure of a business has been the focus of academic research for decades. The overwhelming majority of studies have concentrated on public large enterprises that provide a lot of financial information that can be analysed to establish the determinants/predictors of distress or failure. Two main modelling approaches have emerged: Merton-type structural models (Merton, 1974) that estimate the relationship between default risk and the capital structure of the firm; and reduced-form or accounting-based models that attempt to link observed financial variables with a failure status. Some famous early studies from the latter stream include Altman (1968) and Ohlson (1980). Smaller businesses received far less attention, mainly because of limited information availability. Most of them are not listed which rules out structural models that are based on share price movements, not available for SMEs (although there has been an attempt to apply the Merton model to larger UK SMEs (Lin, 2007)). Therefore, academic research (albeit not extensive) on SME failure modelling focussed on financial statements.

A variety of different classification models have been suggested and applied to samples of SMEs from a range of different countries. Altman and Sabato (2007) investigated the performance of the US SMEs using logistic regression. Sohn and Kim (2013) applied random survival forests to a sample of small businesses in South Korea. Martens et al. (2011) used Support Vector Machines for Flanders. Vallini et al. (2009) found logistic regression to be superior to Multiple Discriminant Analysis (MDA) for Italian SMEs, Ciampi and Gordini (2013) proved that neural networks showed even better performance.

Calabrese and Osmetti (2013) and Calabrese et al. (2016) extended extreme values models (GEV and BGEVA) to predict default of a sample of Italian SMEs, and Andreeva et al. (2016) compared the performance of the UK and Italian small firms.

These studies found that small business failure is associated with profitability, leverage, liquidity, cash flow management, growth and efficiency.

Several studies noted the importance of non-financial information (Altman et al., 2010) found that age and default events in the past were important predictors for financial distress in the UK. Orton et al. (2015) using a large sample of UK SMEs demonstrated that company demographics, derogatory events and information about directors was important.

In mature economies credit risk in lending to SMEs is almost exclusively managed by means of automated credit scoring systems. Credit scoring is a collection of mathematical and statistical models that predict the probability of a borrowers default, using historic data that include the firms characteristics ('demographics'), such as legal form , industry sector, and its financial performance. The main advantage of credit scoring consists in supporting credit expansion while keeping the level of risk under control. There is evidence that it improves access to credit for individuals and SMEs (Berger and Frame, 2007). There are many different classification algorithms that have been applied in credit context to distinguish between defaulting and performing loans, the most common being logistic regression (Thomas et al., 2002). We use this model as a starting point and benchmark for comparison, together with probit regression, which is popular in finance literature.

2.2 Credit contagion and spatial interaction

Different approaches have been proposed in the literature in order to introduce a dependence structure among portfolio exposures. The widely used model assumes that all the exposures are affected by a common factor that represents the economic cycle. This stream is known as the factor model, CreditMetrics (2007) and KMV models are two well known examples of this approach. Unfortunately, most of these models have been proposed for corporates, which are classified in rating classes provided by rating agencies, such as Moody's and Standard & Poor's. Banks and financial regulators have realised that SMEs are a distinct kind of client from corporates, with peculiarities that require specific risk management tools and methodologies (e.g. Altman and Sabato, 2007; Basel Committee on Banking Supervision, 2006 and 2010). Furthermore, few authors have shown that common factors cannot explain all the dependence among defaulted exposures during the last financial crisis (Duffie et al., 2009).

For this reason, several studies have been focused on credit contagion, with an increasing interest after the last financial crisis. Most of these analyses have been focused on corporate defaults. Egloff et al. (2007) model the microeconomic interdependencies among firms using a weighted network and then include them into a macroeconomic factor model. Also Barro and Basso (2010) consider a microecoconomic component to represent the contagion effects, but they include it into a structural approach to model the asset value of the firm. Specifically, they use Monte Carlo simulations first to simulate the networks of firms and then to simulate the possible behaviours on these networks. Cossin and Schellhorn (2007) also use a structural framework, but, instead of a business network, they consider a lending network between firms and apply queueing theory to measure the counterparty risk which can be defined as the risk that the default of a firm's counterparty might affect its own default probability. Analogously to Egloff et al. (2007), Giesecke and Weber (2006) analyse the credit contagion using a network where the edges represent business partner relationships. Under the assumption that firms have the same number of business partner relationships, they introduce the contagion component in a reduced-form model and provide an approximation of the distribution of portfolio losses. The main disadvantage of these approaches is given by the difficulty to implement them because they require information about the network of business partners or the network of lending between firms that are not usually available to financial institutions or policy makers, in particular for small businesses and start-ups.

If we focus our attention on available information for SMEs, a possible solution for including credit contagion in lending decisions could be to use firm locations. The main customers of small businesses are usually located in the same region where the SME operates. Therefore, the probability of default of an SME located in a region can be affected by the performance of the nearby small businesses. Even if several studies (Agarwal and Hauswald, 2010; Degryse and Ongena, 2005; DeYoung et al., 2006; Hauswald and Marquez, 2006; Petersen and Rajan, 2002) have analysed the effects of the distances between small businesses and banks on lending decisions, the role of the distances between SMEs locations on the decisions of granting loans is relatively unexplored in the literature. Barreto and Artes (2013) compute an explanatory variable that represents the spatial interdependence among SMEs and they include it in a logistic regression model to build a scoring model for SMEs. To estimate this covariate, they use the kriging method that takes into account the distance among SMEs and the spatial interdependence associated to a given variable. They apply their proposal to empirical data on Brazilian SMEs and they show that including the spatial component improves the performance of the scoring models. Barro and Basso (2010) generate the locations of firms using an entropy spatial interaction model which takes into account the distances between the regions where the firms are located and the economic sectors of these regions.

3 Methodology

In this paper we suggest to use spatial econometrics in order to model credit contagion in a scoring framework. Furthermore, we propose to use a proximity matrix as a distance matrix in the spatial econometrics approach.

3.1 Spatial models

Let Y_i be the binary dependent variable in a regression model and Y_i^* the latent continuous variable such as

$$Y_i = \begin{cases} 1, & Y_i^* > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

with $i = 1, 2, ..., n$ and n is the sample size. A spatial regression approach (Le Sage and Pace, 2009) provides the following linear model for the latent variable Y_i^*

$$Y^* = \rho W Y^* + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

where X is the $n \times k$ matrix of explanatory variables, ρ is the spatial autocorrelation parameter and $\boldsymbol{\epsilon}$ is the error term. As we use a probit model in this study, we assume that $\boldsymbol{\epsilon}$ is a multivariate normal random variable. W is the weight matrix with main diagonal elements equal to zero and the generic element w_{ij} is different from zero only if the i-th and j-th observations are contiguous. Several approaches have been proposed in the literature to estimate spatial regression models for binary data (see Calabrese and Elkink (2014) for a literature review). In order to estimate the parameters in the equation (3.2), the inverse of the matrix $(I - \rho W)^{-1}$ must be computed.

This calculation can become problematic for a large sample size. The most appropriate method to handle large sample size has been proposed by Klier and McMillen (2008).

Pinkse and Slade (1998) have used the Generalised Method of Moments (GMM) to estimate the parameters of the equation (3.2). Klier and McMillen (2008) have proposed an approximation of Pinkse and Slade's model around $\rho = 0$. They propose their method for a spatial logit model, but we will present it for a spatial probit model as it is the approach used in this study and proposed by Pinkse and Slade (1998). The error term $\boldsymbol{\epsilon}$ are, indeed, distributed as the n-dimensional multivariate normal random variable $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2_{\boldsymbol{\epsilon}} I)$. The variance of the error term is

$$var(\mathbf{e}) = \sigma^2_{\boldsymbol{\epsilon}} \left[(I - \rho W)'(I - \rho W)\right]^{-1}. \tag{3.3}$$

We define

$$D = diag(\boldsymbol{\sigma}_{\mathbf{e}}) \tag{3.4}$$

the diagonal matrix with diagonal elements $\boldsymbol{\sigma}_{\mathbf{e}}$ given by the root square of the diagonal elements in the matrix (3.3). Pinkse and Slade (1998) define the generalised residuals $\mathbf{e}$ as

$$\mathbf{e}(\boldsymbol{\beta}, \rho) = \frac{\phi_n \left[d(\boldsymbol{\beta}, \rho)\right] \left\{\mathbf{y} - \Phi_n[d(\boldsymbol{\beta}, \rho)]\right\}}{\Phi_n \left[d(\boldsymbol{\beta}, \rho)\right] \left\{1 - \Phi_n[d(\boldsymbol{\beta}, \rho)]\right\}}, \tag{3.5}$$

with

$$d(\boldsymbol{\beta}, \rho) = D^{-1}(I - \rho W)^{-1} X\boldsymbol{\beta}. \tag{3.6}$$

Analogously to an independent and identically distributed probit model, the aim of the spatial probit is to estimate the following function

$$\mathbf{P} = P\{Y = 1/d(\boldsymbol{\beta}, \rho)\} = \Phi\left[d(\boldsymbol{\beta}, \rho)\right] \tag{3.7}$$

where $d(\boldsymbol{\beta}', \rho)$ is defined in equation (3.6).

The objective function of the GMM approach is

$$(\hat{\boldsymbol{\beta}}', \hat{\rho})' = \arg \min_{(\boldsymbol{\beta}', \rho)'} \mathbf{e}'(\boldsymbol{\beta}', \rho)' Z (Z'Z)^{-1} Z' \mathbf{e}(\boldsymbol{\beta}', \rho)',$$

where Z is a matrix of instrument variables and D is defined in equation (3.4).

Klier and McMillen (2008) have suggested to use a nonlinear two-stage least squares method. Given some initial values $(\boldsymbol{\beta}_0', \rho_0)'$, the generalised residuals $\mathbf{e}_0$ are computed using the equation (3.5). The gradient terms are

$$\begin{aligned} G_{\boldsymbol{\beta} i} &= \frac{\partial P_i}{\partial \boldsymbol{\beta}} = \hat{P}_i(1-\hat{P}_i)\mathbf{t}_i \\ G_{\rho i} &= \frac{\partial P_i}{\partial \rho} = \hat{P}_i(1-\hat{P}_i)\left[h_i - \frac{d_i}{\sigma_{ei}^2}\Upsilon_{ii}\right], \end{aligned} \tag{3.8}$$

where $\mathbf{t}_i$ is the i-th row vector of the matrix $T = D^{-1}(I-\rho W)^{-1}X$, h_i is the i-th element of the vector $\mathbf{h} = (I-\rho W)^{-1}W\mathbf{q}$, d_i is the i-th element of the vector $\mathbf{d}$ defined in equation (3.6) and Υ_{ii} is the i-th element of the diagonal of the matrix $\Upsilon = (I-\rho W)^{-1}W(I-\rho W)^{-1}(I-\rho W)^{-1}$.

The two stages of Klier and McMillen's approach are:

I stage: the gradient terms $G_{\boldsymbol{\beta}}$ and G_ρ are regressed on Z and compute the predicted values $\hat{G}_{\boldsymbol{\beta}}$ and $\hat{G}_\rho$;

II stage: $\mathbf{e}_0 + G_{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}_0$ are regressed on $\hat{G}_{\boldsymbol{\beta}}$ and $\hat{G}_\rho$. The coefficients obtained from this regression are the estimated values of $\boldsymbol{\beta}$ and ρ.

An important aspect of spatial econometrics is to build a coherent distance matrix W. To do this, we use the proximity matrices adopted in data mining.

3.2 Proximity matrices

Orton et al (2015) report marked differences in SME default rates across different UK regions, so it is reasonable to assume that SMEs close to each other geographically will experience similar economic conditions and this will be reflected in their performance. Our first dependence model explores this kind of interdependence with businesses in the same or adjacent postcodes marked as connected, which is indicated by 1 in the W matrix. Companies in non-adjacent postcodes are marked by 0.

Yet the dependence may not necessarily be confined to the spatial element, and companies in the same industrial sector or of the same legal form may also be connected and exhibit similar behaviour. We explore this kind of demographic dependence by creating a W matrix based on two proximity measures:

- proximity arising from similarity in formal descriptors of the type of the company, measured by Jaccard similarity index
- proximity arising from the above plus distance in Age, measured by Gower's similarity.

The former is a commonly used measure for binary categorical variables, and the latter can handle variables of different types (Gower and Legendre, 1986). We have separated Age because several authors commented on its importance (Altman et al., 2010; Orton et al., 2015).

Let $s(x, y)$ be a measure of similarity for observations x and y. Jaccard index is defined as:

$$s_{Jac}(x, y) = \frac{\sum_{j=1}^{v} \delta_{x,y}^{j}}{v}, \tag{3.9}$$

where v is the number of variables, for which either x or y is non-zero,

$$\delta_{x,y}^{j} = \begin{cases} 1, & \text{if } x_j = y_j \text{ for variable } j \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

Gower's similarity measure is computed as:

$$s_G(x, y) = \frac{\sum_{j=1}^{v} \delta_{x,y}^{j} d_{x,y}^{j}}{\sum_{j=1}^{v} \delta_{x,y}^{j}}, \tag{3.11}$$

$\delta^j_{x,y} = 1$ for variable j, if it is nominal, ordinal, interval or ratio;
or for asymmetric nominal

$$\delta^j_{x,y} = \begin{cases} 1, & \text{if } x_j \text{ or } y_j \text{ is present} \\ 0, & \text{if both } x_j \text{ or } y_j \text{ are absent;} \end{cases} \tag{3.12}$$

$d^j_{x,y} = 1$ - $|x_j - y_j|$, for ordinal, interval, ratio variable;
or for nominal

$$\delta^j_{x,y} = \begin{cases} 1, & \text{if } x_j = y_j \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

To model the interaction between spatial and demographic components, 1s in the adjacency matrix W are replaced by either Jaccard or Gower distances, which vary between 0 and 1, indicating different levels of dependence.

4 Empirical analysis

4.1 Data

The data for analysis comes from a large anonymised database from a credit bureau that covers more than two mln enterprises, all enterprises that borrow from the financial institutions. We apply the definition of SME proposed by the European Commission: the annual turnover is lower than 50 million Euro and the number of employees should not exceed 250.

A year of 2007 just before the crisis was used to observe the credit performance. There are 92 potential predictors describing the company demographics, information about directors, payment and public records, accounting ratios and trends. The default indicator is supplied by the data provider and includes dissolved companies, as well as those in liquidation, receivership or under administration.

London area was selected as the most suitable for our analysis with its high concentration of small businesses. The current research considers established small businesses that are at least 3 years old. The reason for this is that younger enterprises, and in particular start-ups, show very different behaviour and significantly higher default rates (Orton et al., 2015). A random sample of 10 percent was selected from all London non-start-up companies, with 27,533 companies forming the sample for analysis.

To be consistent with the existing credit scoring methodology, coarse-classification or binning is used to transform the predictors (Thomas et al., 2002). Numeric variables are divided into 10 classes that are subsequently grouped together if the default rates in adjacent categories are close. For categorical variables, small categories are banded together to improve the robustness of the model. This procedure also allows to cope with outliers and missing values, the latter become a separate category. The final coarse-classes can be entered into the model as binary dummy variables, or a more widely spread alternative, which we follow in this paper, would be to transform them into Weights of Evidence (WoE):

$$WoE_i = ln\left(\frac{h_i/d_i}{H/D}\right) = ln\left(\frac{h_i\,D}{d_i\,H}\right), \tag{4.1}$$

where h_i is the number of healthy companies in category i of a variable, d_i is the number of defaults in category i, H is the total number of healthy companies in the sample, D is the total number of defaults.

WoE transformation is popular and widely used in the credit scoring industry (Crook et al., 2007; Thomas, 2009). Lin et al. (2012) reported improved predictive performance when using WoE in small business distress context.

We consider the following types of dependency between the enterprises:

1. Spatial dependence based on geographic proximity. First two letters of the postcode are available from the data, the full postcode could not be provided because of confidentiality reasons. Companies within the

same postcode and in adjacent postcodes are considered as connected and marked with 1, all other companies are marked as 0.[1]

2. Demographic dependence based on the following categorical variables and measured as proximity distance by Jaccard index:

    - Legal Form
    - No of Employees (categorised) - as the measure of company size
    - Company is Subsidiary
    - Ultimate Parent Company - derogatory details
    - Parent Company - derogatory details
    - Industry - first digit of SIC 1992 code
    - Company has subsidiary.

3. Demographic dependence based on variables in (2) and Age in months from the date of incorporation, measured by Gower similarity coefficient.

The variation in the number of neighbours could generate heteroscedasticity, so we normalised W such as $w_{ij}/(\sum_j w_{ij})$ for $i, j = 1, 2, \ldots, n$ following the literature (LeSage and Pace, 2009).

4.2 Empirical results

We apply a multicollinearity analysis to the 92 available predictors. The explanatory variables with a Variance Inflation Factor lower than 5 are reported in Table 1. Afterwards, we randomly split the 27,533 SMEs into two groups. In this section we analyse the estimates that we obtain on 90% of the firms (our training sample). We use the spatial probit model suggested

[1]According to the spatial econometrics literature (LeSage and Pace, 2009), a firm is not connected with itself, so the elements on the main diagonal are equal to 0.

by Klier and McMillen (2008) presented in Section (3.1) and implemented in the R package 'McSpatial'. The parameters $\boldsymbol{\beta}$ and ρ of the model (3.2) are estimated for three different weight matrices W: geo, Jaccard and Gower. Table 1 shows the results for the geographical contiguity matrix, Table 2 for Jaccard matrix defined in equation (3.9) and Table 3 for Gower matrix presented in equation (3.11). Each table shows different groups of explanatory variables because we remove the covariates used to build Jaccard and Gower matrices in order to avoid endogeneity (LeSage and Pace, 2008). The upper part of each table shows the results for the independent and identically distributed probit model, the lower those for the spatial probit model.

We follow the decision making procedure used in the industry. We estimate the WoE defined in equation (4.1). We consider the dependent variable $Y = 1$ for defaulted SMEs. Therefore, the signs of the estimates should be negative. As Table 1, 2 and 3 show, most of the estimates are negative. Few variables, i.e. 'number of employees', 'number of consolidated accounts' and 'number of days between filed accounts date and date recorded at companies house' show positive signs in most of the tables. The result obtained for consolidated accounts is consistent because this variable is not significant. Instead, 'number of employees' and 'number of days between filed accounts date and date recorded at companies house' might be correlated with the remaining variables even though we have checked for multicollinearity.

The last rows in Table 1, 2 and 3 report the results for the autocorrelation parameter ρ. Table 1 shows that ρ is not significant if we use the geographical adjacency matrix. This means that at two-digit postcode level the spatial proximity of SMEs is not relevant to explain the interdependence between distressed firms. We may obtain this result as the London area is limited and affected by similar economic conditions. In order to better understand the mechanism of default propagation, it is more important to take into account SMEs' characteristics such as industry sector, legal form, number of employees and presence of subsidiary. In this context, the intensity of interdependence between SMEs' defaults is high, where Jaccard matrix shows a higher estimate of ρ (0.737) than the one for Gower matrix (0.637). The level of significance of the parameter ρ for the Jaccard matrix is lower than those for Gower matrix. This result does not support our expectations given the importance of Age in previous studies. Further research will investigate the role of Age in more detail with exploration of different proximity measures.

McMillen (1992) notes that ignoring the interdependence and, indeed, assuming independent observations can cause the parameter estimates to be inconsistent, as the error terms are heteroskedastic. The identification of SMEs' risk drivers is crucial in decision making processes for financial institutions and policy makers. The results in Table 1, 2 and 3 show the effects of the independence assumption on the parameter estimates. For example, including or ignoring the contagion effects between SMEs highly affects the estimate of the intercept, as Table 2 and 3 show. The independence assumption also affects the standard deviation and, indeed, the level of significance of the independent variables. If we consider the linkages between the spatial locations of SMEs, the variables 'number of employees' and 'type of accounts' become highly significant, coherent with Andreeva et al. (2016). Instead, if we use Jaccard or Gower matrix,'number of previous searches' turns into an important risk driver, in line with Orton et al. (2015). The two proximity matrices generate also different effects on the levels of significance. 'The number of unsatisfied mortgages' becomes highly significant using the Jaccard matrix, whilst 'days between filed accounts date and date recorded at companies house' and 'types of accounts' become highly significant with the Gower matrix.

4.3 Impact on decision making

There are two important implications of our analysis for decision-making. The first one concerns the interpretation of parameter estimates in the risk model as discussed in the previous section. Knowledge of significant risk drivers of the default provides the information to policy-makers on the early signs of default, in addition our models give insights on how the estimated defaults may be connected. Such information can be used for planning interventions or campaigns to support the failing businesses. This is equally important for lenders. Yet for the latter the second vital aspect of the decision-making is the ability to accurately predict which companies are more likely to fail in order to avoid bad debt losses. Accurate estimates of the probability of default (PD) are also required to calculate the amount of regulatory capital the banks are required to hold as the buffer against the bad debt (Basel Committee, 2006, 2010).

In this section we concentrate on the predictive accuracy of the models using the measures most commonly used in predictive modelling and by the credit industry (Thomas, 2002). Area under the ROC curve (AUC) presents the true positive rate (TPR) or the proportion of correctly predicted defaults against the false positive rate (FPR) or the proportion of incorrectly predicted defaults over all potential threshold values. It was shown by Hanley and McNeil (1982) that conceptually AUC corresponds to the Wilcoxon or Mann-Whitney statistic, which estimates the probability that a predicted PD of a randomly selected defaulted business is higher than or equal to that of a healthy business. We also report separately FPR as the proportion of wrongly classified healthy companies from all healthy companies, and false negative rate (FNR) as the proportion of incorrectly classified defaults from all defaults. In order to decide whether the company should be classified as default/ non-default, we use a cut-off corresponding to the average of the mean predicted default probabilities for failed and healthy companies.

Although AUC is a popular measure in data mining and credit scoring, Hand (2009) argued that in certain applications it is necessary to take into account the costs of misclassification as these might differ between the two types of error. In lending environment the cost of misclassifying a default is higher than misclassifying a healthy company since granting credit and experiencing a loss is more costly than the lost opportunity of turning away a healthy credit applicant. Therefore, we also report H measure (Hand, 2009) that incorporates the cost imbalance. For both AUC and H, higher values correspond to better prediction.

Two other popular measures - Mean Absolute Error (MAE) and Mean Square Error (MSE) - are based on the difference between observed and predicted values. Given the argument above about misclassification costs, these measures are reported for defaults only in order to show separately the predictive accuracy for failing businesses.

Predictive accuracy is summarised in Table 4. In order to ensure that our predictive accuracy results are not based just on one test sample, we perform 100-fold cross-validation, which is a standard approach in data mining and predictive modelling. We randomly split the dataset into training (90%) and test (10%) sets, estimate the models on the training set and apply them to the test one. We repeat the procedure 100 times. Table 4 reports average measures with standard deviations given in brackets. In previous section 4.2. the estimates are reported from the training sample obtained from the first split.

For spatial model with geo-matrix there is very little difference between all measures reported. Nevertheless, one can argue that spatial model provides similar level of predictive accuracy on top of providing additional insights into the spatial connectedness of SMEs. In contrast, the models with Jaccard and Gower matrices demonstrate much more encouraging results. Whilst the values for AUC and H appear marginal, the paired t-test indicates that they are different with the Probit model the best. Yet, for the MAE, MSE and FNR, there is a marked improvement in the measures that concentrate on default. We would like to reiterate that erroneously classifying defaulting company as a healthy one is more costly as compared to the other way round.

We present the histograms of MAE on defaulted SMEs for the spatial model using Gower matrix in Figure 1 and for the probit model in Figure 2.

We do not show the histograms of the MSE for these models and the MAE as they show similar behaviours to those in Figure 1 and Figure 2. From these plots, we observe that the variability of these error measures is lower for the probit model only because their distributions are highly skewed towards high values of the errors.

Table 5 provides some additional information on prediction for defaults, by reporting five-point summaries for the estimated PD distributions for problematic companies, also obtained from 100-fold cross-validation. Whilst again there is little difference for spatial geo-model, both Jaccard and Gower show a pronounced shift of distributions towards the higher range, closer to 1, which indicates more accurate predictions.

5 Discussion and further research

This paper presented the first attempt to investigate the nature of interdependence of SME defaults. A starting point of this investigation is the spatial dependence framework developed in spatial econometrics (Le Sage and Pace, 2009). This framework is combined with elements of standard methodology of credit risk assessment.

Yet we do not find the evidence of geographic dependence. The spatial model with contiguity matrix for SMEs in adjacent postcodes does not have a statistically significant estimate for the autocorrelation parameter. Nevertheless, the innovation and the main contribution of this paper is to extend to spatial dependence framework beyond geography. Adding the demographic dependence as measured by two proximity measures - Jaccard and Gower - showed statistically significant autocorrelation coefficients, change in parameter estimates of predictor variables, and improved predictive accuracy for defaulting companies.

This implies that policy-makers when deciding on areas of support for SME development and lenders when assessing risk of SME default should take into account the correlated nature of defaults of business with similar demographic characteristics, such as legal form, industry sector and number of employees. This is important, especially in times of economic shocks when the default rate can rapidly rise. Banking regulations (Basel Committee, 2006, 2010) stipulate that for conservative estimates of bad debt losses, the lenders are required to stress-test the models under adverse conditions. Failure to incorporate the interdependence between defaults will lead to underestimation of losses.

By investigating the nature of the default dependencies, this paper provided the first step towards the framework for incorporating dependence into credit risk assessment. Further work will involve exploration of different proximity measures and different combination of variables defining the dependence matrix. An important extension includes stress-testing framework, which will incorporate macroeconomic variables.

6 References

Agarwal, S., Hauswald, R. (2010) Distance and private information in lending. The Review of Financial Studies 23(7), 2757-2788.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589-609.

Altman, E., Sabato, G. (2007). Modeling credit risk for SMEs: Evidence from the US market. ABACUS, 43(3), 332-357.

Altman, E., Sabato, G., Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. The Journal of Credit Risk, 6, 1-33.

Andreeva, G., Calabrese, R., Osmetti, A. S. (2016) A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. European Journal of Operational Research 249 (2), 506-516.

Barreto, G., Artes, F. (2013) Spatial correlation in credit risk and its improvement in credit scoring. Insper Working Paper.

Barro, D., Basso, A. (2010) Credit contagion in a network of firms with spatial interaction. European Journal of Operational Research, 205(2), 459-468.

Basel Committee on Banking Supervision (2006). International Convergence of Capital Measurement and Capital Standards A Revised Framework Comprehensive Version.

Basel Committee on Banking Supervision (2010), Basel III: International framework for liquidity risk measurement, standards and monitoring.

Battiston, S., Delli Gatti, D., Gallegatic, M., Greenwaldd, B., Stiglitz, J. E. (2012) Default cascades: When does risk diversification increase stability? Journal of Financial Stability, 8, 138-149.

Berger, A. N., Udell, G. F. (2002) Small Business Credit Availability and Relationship Lending: The Importance of Bank Organisational Structure. The Economic Journal 112 (477), 32-53.

Calabrese, R., Osmetti, S. A. (2013). Modelling SME loan defaults as rare events: the Generalized Extreme Value regression model. Journal of Applied Statistics, 40(6), 1172-1188.

Calabrese, R., Elkink, J. (2014) Estimators of binary spatial autoregressive models: A Monte Carlo study. Journal of Regional Science, 54(4), 664-687.

Calabrese, R., Marra, G., Osmetti, S. A. (2016) Bankruptcy Prediction of Small and Medium Enterprises Using a Flexible Binary Generalized Extreme Value Model. 67 (4), 604615.

Ciampi F., Gordini N. (2013) Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises. Journal of Small Business Management, 51(1), 23-45.

Cossin, D., Schellhorn, H., 2007. Credit risk in a network economy. Management Science, 1604-1617.

Cravo, T. A., Becker B., Gourlay, A. (2015) Regional Growth and SMEs in Brazil: A Spatial Panel Approach. Regional Studies, 49(12), 1995-2016

CreditMetrics, 2007. CreditMetrics (TM) technical document, riskMetrics group. www.riskmetrics.com

Crook, J. N., Edelman, D., Thomas, L. C. (2007) Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183, 1447-1465.

Delli Gatti, D., Gallegati, M., Greenwald, B. C., Russo, A., Stiglitz, J. E. (2009) Business Fluctuations and bankruptcy avalanches in an evolving network economy. Journal of Economic Interactions and Coordination, 4, 195-212.

Degryse, H., and S. Ongena. 2005. Distance, Lending Relationships, and Competition. Journal of Finance 60, 231-66.

DeYoung, R., Glennon, D., Nigro, P. (2006) BorrowerLender Distance, Credit Scoring, and the Performance of Small Business Loans. FDIC Center for Financial Research Working Paper No. 2006-04.

Duffie, D., Eckner, A., Horel. G., Saita, L. (2009) Frailty correlated defaults. The Journal of Finance, 64(5),

Egloff, D., Leippold, M., Vanini, P. (2007) A simple model of credit contagion. Journal of Banking and Finance, 31, 2475-2492.

Fernandes, G. B., Artes, R. (2016) Spatial dependence in credit risk and its improvement in credit scoring. European Journal of Operational Research, 249 (2), 517-524.

Ferretti, V., Montibeller, G. (2016) Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. Decision Support Systems, 84, 41-52.

Giesecke, K., Weber, S. (2006) Credit contagion and aggregate losses. Journal of Economic Dynamics & Control 30, 741-767.

Gower, J. C., Legendre, P. Metric and Euclidean properties of dissimilarity coefficients. Journal of classification. 1986, 5-48.

Gnyawali, D. R., Byung-Jin, R. (2009) Co-opetition and Technological

Innovation in Small and Medium-Sized Enterprises: A Multilevel Conceptual Model. Journal of Small Business Management 2009 47(3), 308-330.

Grunert, J., Norden, L. (2012) Bargaining power an infomation in SME lending. Small Business Economics, 39 (2), 401-417.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103-123.

Hanley, J. A., McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Kazemi, R., Mosleh, A. (2012) Improving Default Risk Prediction Using Bayesian Model Uncertainty Techniques. Risk Analysis, 32(11), 1888-1900.

Klier, T., McMillen, D. P. (2008) Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples. Journal of Business & Economic Statistics 26(4), 460-471.

LeSage, J., Pace, R. K. (2009) Introduction to Spatial Econometrics. CRC Press.

Lin, S. M. (2007) SMEs Credit Risk Modelling for Internal Rating Based Approach in Banking Implementation of Basel II Requirement. PhD thesis. University of Edinburgh.

Lin, S. M., Ansell, J., Andreeva, G. (2012) Predicting default of a small business using different definitions of financial distress. Journal of the Operational Research Society, 63, 539-548.

Loebecke, C., Van Fenema, P., Powell, P. (1999) Coopetition and Knowledge Transfer. ACM SIGMIS Database - Special issue on information systems: current issues and future changes, 30 (2), 14-25.

Martens, D., Vanhoutte, C., Winne, S. D., Baesens, B., Sels, L., Mues, C. (2011) Identifying Financially Successful Start-Up Profiles with Data Mining. Expert Systems with Applications, 38, 5794-5800.

McMillen, D. P. (1992) Probit with Spatial Autocorrelation, Journal of Regional Science, 32(3), 335-348.

Merton, R. (1974) On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance, 29, 449-470.

Neu, P., Khun, R. (2004) Credit risk enhancement in a network of interdependent firms. Physics A, 342, 639-655.

Ohlson, J. A. (1980) Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18, 109-131.

Orton, P., Ansell, J., and Andreeva, G. (2015) Exploring the performance of small- and medium-sized enterprises through the credit crunch. Journal of the Operational Research Society, 66, 657-663.

Petersen, M., Rajan, R. (2002) Does Distance Still Matter? The Information Revolution in Small Business Lending. Journal of Finance 57, 253-370.

Pinkse, J., Slade, M. (1998) Contracting in Space: An Application of Spatial Statistics to Discrete Choice Models. Journal of Econometrics 85, 125-154.

Sohn, S. Y. and Kim, Y. S. (2013) Behavioral credit scoring model for technology-based firms that considers uncertain financial ratios obtained from relationship banking. Small Business Economics, 41(4), 931-943.

Thomas, L., Edelman, D., and Crook, J. C. (2002) Credit Scoring and Its Applications. Society for Industrial and Applied Mathematics, Philadelphia.

Thomas, L. (2009) Consumer Credit Models. Oxford University Press: Oxford.

Vallini, C. F., Ciampi, F., Gordini, N., Benvenuti, M. (2009) Are Credit Scoring Models Able To Predict Small Enterprise Default? Statistical Evidence from Italian Small Enterprises. International Journal of Business & Economics, 8(1), 3-18.

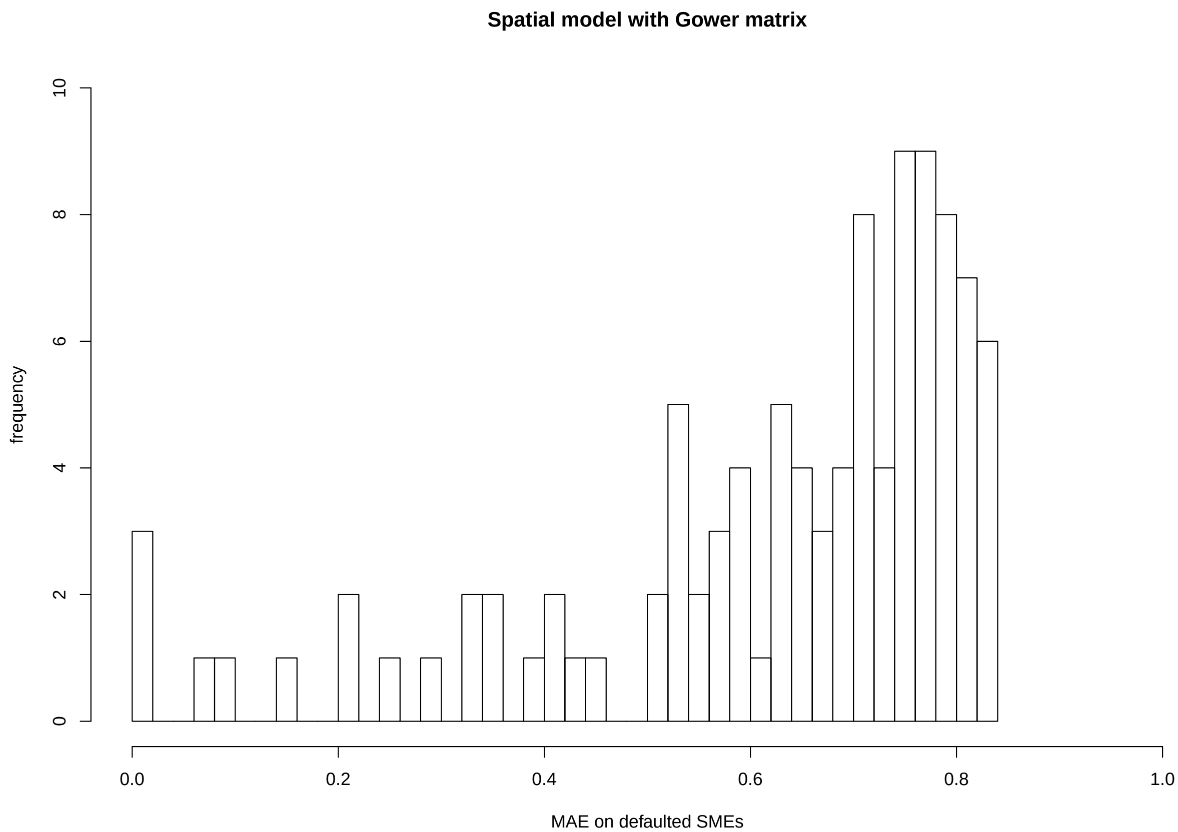

Figure 1: The histogram of MAE on defaulted SMEs for the spatial model with Gower matrix

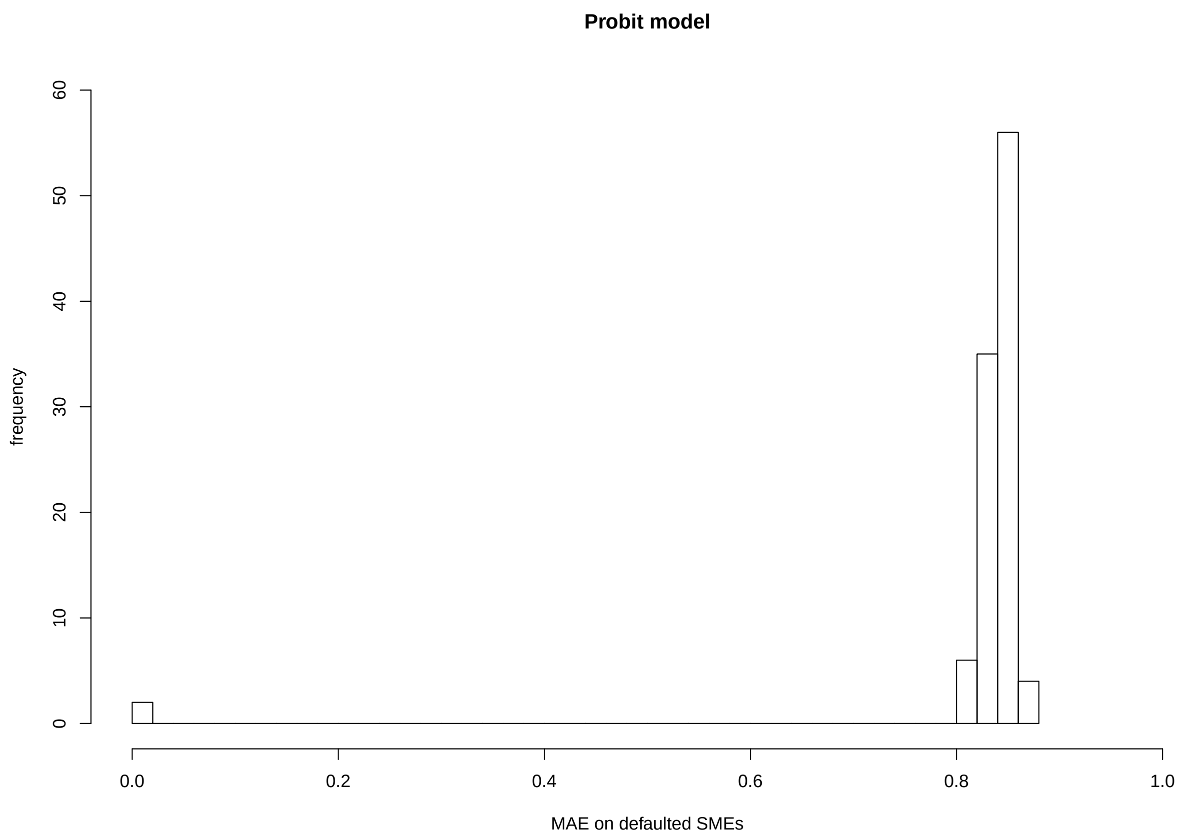

Figure 2: The histogram of MAE on defaulted SMEs for the probit model

PROBIT MODEL				
Variables	*Estimates*	*Std. Error*	*z-value*	$Pr(>\|z\|)$
Intercept	-1.719	0.0172	-99.905	0.000
Legal Form	-0.762	0.117	-6.542	0.000
Age	-0.160	0.043	-3.741	0.000
No of Employees	0.541	0.406	1.333	0.182
Current Directors to Previous Directors	-0.283	0.083	-3.225	0.001
Worst DBT in Last 12 Months	-0.159	0.056	-2.851	0.004
No of Previous Searches	-0.374	0.157	-2.384	0.017
Time since Last Derogatory Item	-0.192	0.028	-6.732	0.000
No of Unsatisfied Mortgages and Charges	-0.449	0.173	-2.590	0.009
Lateness of Accounts	-0.396	0.018	-22.332	0.000
Days btw Filed Accounts Date and Date Recorded at Companies House	0.109	0.046	2.379	0.017
Consolidated Accounts	0.031	0.227	0.135	0.892
Type of Accounts	-0.159	0.152	-1.044	0.296
Time Since Last Annual Return	-0.328	0.022	-15.081	0.000
Current Liabilities	-0.529	0.079	-6.666	0.000
Percentage Change in Total Assets	-0.270	0.033	-8.131	0.000

SPATIAL PROBIT MODEL				
Variables	*Estimates*	*Std. Error*	*z-value*	$Pr(>\|z\|)$
Intercept	-1.838	0.455	-4.041	0.000
Legal Form	-0.790	0.094	-8.415	0.000
Age	-0.164	0.023	-7.014	0.000
No of Employees	0.516	0.226	2.280	0.023
Current Directors to Previous Directors	-0.290	0.054	-5.319	0.000
Worst DBT in Last 12 Months	-0.160	0.032	-4.968	0.000
No of Previous Searches	-0.369	0.092	-3.987	0.000
Time since Last Derogatory Item	-0.193	0.026	-7.294	0.000
No of Unsatisfied Mortgages and Charges	-0.455	0.086	-5.259	0.000
Lateness of Accounts	-0.398	0.013	-29.627	0.000
Days btw Filed Accounts Date and Date Recorded at Companies House	0.111	0.031	3.606	0.000
Consolidated Accounts	-0.028	0.152	-0.183	0.855
Type of Accounts	-0.152	0.085	-1.785	0.074
Time since Last Annual Return	-0.328	0.016	-20.171	0.000
Current Liabilities	-0.534	0.048	-11.032	0.000
Percentage Change in Total Assets	-0.269	0.020	-13.254	0.000
WXB	-0.059	0.229	-0.256	0.798

Table 1: Parameter estimates for a probit model and a spatial probit model with weight matrix W given by the geographic adjacency matrix.

PROBIT MODEL				
Variables	*Estimates*	*Std. Error*	*z-value*	$Pr(>\|z\|)$
Intercept	-1.691	0.016	-104.382	0.000
Age	-0.167	0.042	-3.924	0.000
Current Directors to Previous Directors	-0.320	0.085	-3.748	0.000
Worst DBT in Last 12 Months	-0.162	0.056	-2.901	0.003
No of Previous Searches	-0.343	0.157	-2.190	0.028
Time since Last Derogatory Item	-0.190	0.029	-6.664	0.000
No of Unsatisfied Mortgages and Charges	-0.426	0.173	-2.456	0.014
Lateness of Accounts	-0.388	0.018	-22.027	0.000
Days btw Filed Accounts Date and Date Recorded at Companies House	0.107	0.045	2.362	0.018
Consolidated Accounts	-0.097	0.216	-0.451	0.652
Type of Accounts	-0.121	0.136	-0.887	0.375
Time since Last Annual Return	-0.331	0.021	-15.355	0.000
Current Liabilities	-0.471	0.079	-5.981	0.000
Percentage Change in Total Assets	-0.282	0.033	-8.516	0.000

SPATIAL PROBIT MODEL				
Variables	*Estimates*	*Std. Error*	*z-value*	$Pr(>\|z\|)$
Intercept	-0.273	0.599	-0.457	0.648
Age	-0.172	0.023	-7.522	0.000
Current Directors to Previous Directors	-0.306	0.052	-5.902	0.000
Worst DBT in Last 12 Months	-0.163	0.032	-5.099	0.000
No of Previous Searches	-0.327	0.091	-3.574	0.000
Time since Last Derogatory Item	-0.190	0.026	-7.219	0.000
No of Unsatisfied Mortgages and Charges	-0.434	0.086	-5.060	0.000
Lateness of Accounts	-0.385	0.014	-28.394	0.000
Days btw Filed Accounts Date and Date Recorded at Companies House	0.102	0.031	3.334	0.000
Consolidated Accounts	0.001	0.168	0.006	0.994
Type of Accounts	-0.089	0.074	-1.218	0.223
Time since Last Annual Return	-0.324	0.016	-20.260	0.000
Current Liabilities	-0.490	0.048	-10.305	0.000
Percentage Change in Total Assets	-0.278	0.020	-13.848	0.000
WXB	0.737	0.312	2.359	0.018

Table 2: Parameter estimates for a probit model and a spatial probit model with weight matrix W given by Jaccard matrix.

PROBIT MODEL				
Variables	*Estimates*	*Std. Error*	*z-value*	$Pr(>\|z\|)$
Intercept	-1.693	0.016	-104.992	0.000
Current Directors to Previous Directors	-0.326	0.085	-3.824	0.000
Worst DBT in Last 12 Months	-0.168	0.056	-3.025	0.002
Number of Previous Searches	-0.326	0.157	-2.078	0.038
Time since Last Derogatory Item	-0.197	0.028	-6.960	0.000
Number of Unsatisfied Mortgages and Charges	-0.483	0.173	-2.797	0.005
Lateness of Accounts	-0.400	0.017	-23.161	0.000
Days btw Filed Accounts Date and Date Recorded at Companies House	0.110	0.045	2.423	0.015
Consolidated Accounts	-0.080	0.212	-0.375	0.707
Type of Accounts	-0.197	0.135	-1.463	0.143
Time since Last Annual Return	-0.332	0.021	-15.436	0.000
Current Liabilities	-0.482	0.079	-6.130	0.000
Percentage Change in Total Assets	-0.294	0.033	-8.935	0.000
SPATIAL PROBIT MODEL				
Variables	*Estimates*	*Std. Error*	*z-value*	$Pr(>\|z\|)$
Intercept	-0.469	0.623	-0.752	0.451
Current Directors to Previous Directors	-0.311	0.051	-6.030	0.000
Worst DBT in Last 12 Months	-0.170	0.032	-5.303	0.000
Number of Previous Searches	-0.306	0.091	-3.348	0.000
Time since Last Derogatory Item	-0.196	0.026	-7.509	0.000
Number of Unsatisfied Mortgages and Charges	-0.488	0.086	-5.678	0.000
Lateness of Accounts	-0.395	0.013	-29.695	0.000
Days btw Filed Accounts Date and Date Recorded At Companies House	0.104	0.031	3.394	0.000
Consolidated Accounts	0.045	0.161	0.277	0.782
Type of Accounts	-0.167	0.073	-2.297	0.022
Time since Last Annual Return	-0.325	0.016	-20.324	0.000
Current Liabilities	-0.494	0.047	-10.436	0.000
Percentage Change in Total Assets	-0.290	0.020	-14.573	0.000
WXB	0.637	0.326	1.955	0.050

Table 3: Parameter estimates for a probit model and a spatial probit model with weight matrix W given by Gower matrix.

	Spatial model with contiguity matrix	*Logit model*	*Probit model*	*t-statistic SpatialvLogit*	*t-statistic SpatialvProbit*
AUC	0.802(0.019)	0.802(0.019)	0.802(0.019)	-0.993(0.323)	-1.226(0.223)
H	0.289(0.046)	0.290(0.046)	0.289(0.046)	-1.225(0.223)	-0.475(0.636)
MAE^+	0.835(0.013)	0.830(0.014)	0.836(0.013)	28.129(0.000)	-8.146(0.000)
MSE^+	0.724(0.020)	0.720(0.021)	0.725(0.019)	17.861(0.000)	-7.252(0.000)
FNR	0.525(0.034)	0.545(0.034)	0.524(0.034)	-17.204(0.000)	1.431(0.156)
FPR	0.077(0.007)	0.067(0.006)	0.077(0.007)	49.976(0.000)	3.085(0.030)
	Spatial model with Jaccard matrix	*Logit model*	*Probit model*	*t-statistic SpatialvLogit*	*t-statistic SpatialvProbit*
AUC	0.794(0.020)	0.795(0.020)	0.796(0.020)	-4.835(0.000)	-5.565(0.000)
H	0.273(0.045)	0.277(0.045)	0.276(0.046)	-3.889(0.000)	-3.802(0.000)
MAE^+	0.493(0.228)	0.835(0.014)	0.840(0.012)	-15.008(0.000)	-15.008(0.000)
MSE^+	0.338(0.205)	0.726(0.020)	0.731(0.019)	-18.948(0.000)	-19.180(0.000)
FNR	0.386(0.102)	0.543(0.034)	0.524(0.033)	-16.078(0.000)	-14.308(0.000)
FPR	0.189(0.123)	0.069(0.007)	0.078(0.007)	9.863(0.000)	9.029(0.000)
	Spatial model with Gower matrix	*Logit model*	*Probit model*	*t-statistic SpatialvLogit*	*t-statistic SpatialvProbit*
AUC	0.793(0.020)	0.795(0.020)	0.795(0.020)	-5.104(0.000)	-5.104(0.000)
H	0.271(0.046)	0.274(0.046)	0.274(0.046)	-4.267(0.000)	-4.690(0.000)
MAE^+	0.628(0.203)	0.835(0.013)	0.840(0.012)	-10.224(0.000)	-10.485(0.000)
MSE^+	0.477(0.198)	0.727(0.020)	0.731	-12.631(0.000)	-12.631(0.000)
FNR	0.447(0.080)	0.547(0.030)	0.524(0.033)	-13.287(0.000)	-10.613(0.000)
FPR	0.137(0.079)	0.068(0.006)	0.077(0.007)	8.688(0.000)	-7.500(0.000)

Table 4: Mean of forecasting accuracy measures for 100 out-of-sample sets (standard deviation in brackets). The results of paired t-test: test statistics (p-value in brackets). MAE^+ and MSE^+ are errors computed only on defaulted SMEs.

	Spatial model with contiguity matrix	*Logit model*	*Probit model*
Mean	0.164(0.013)	0.169(0.014)	0.163(0.013)
First Quartile	0.036(0.005)	0.034(0.005)	0.036(0.005)
Median	0.093(0.017)	0.086(0.016)	0.093(0.017)
Third Quantile	0.261(0.029)	0.267(0.036)	0.260(0.029)
Max	0.648(0.045)	0.709(0.048)	0.644(0.044)
	Spatial model with Jaccard matrix	*Logit model*	*Probit model*
Mean	0.506(0.228)	0.165(0.013)	0.159(0.012)
First Quartile	0.334(0.271)	0.033(0.004)	0.034(0.005)
Median	0.471(0.264)	0.084(0.017)	0.091(0.018)
Third Quantile	0.679(0.207)	0.262(0.033)	0.255(0.027)
Max	0.898(0.096)	0.689(0.051)	0.629(0.045)
	Spatial model with Gower matrix	*Logit model*	*Probit model*
Mean	0.371(0.203)	0.164(0.013)	0.159(0.012)
First Quartile	0.200(0.222)	0.033(0.004)	0.034(0.005)
Median	0.316(0.234)	0.083(0.016)	0.090(0.017)
Third Quantile	0.529(0.207)	0.261(0.033)	0.252(0.026)
Max	0.829(0.114)	0.696(0.052)	0.636(0.047)

Table 5: The mean of some descriptive statistics of the predicted values for defaults on 100 out-of-sample sets (standard deviation in brackets).